Check for updates

Received: 12 October 2024 | Revised: 7 March 2025 | Accepted: 7 March 2025

DOI: 10.1002/jac5.70038

CPRR: CLINICAL PHARMACY RESEARCH REPORT

jaccp Journal of the American College of Clinical Pharmacy

Accuracy and reproducibility of ChatGPT responses to real-world drug information questions

Shikha Khatri Pharm.D.[1] | Anthony Sengul Pharm.D.[1] | Jungyeon Moon Pharm.D., Ph.D.[1,2] | Cynthia A. Jackevicius Pharm.D., M.Sc., FCCP[1,2,3,4]

[1]Western University of Health Sciences, Pomona, California, USA

[2]VA Greater Los Angeles Healthcare System, Los Angeles, California, USA

[3]ICES, Toronto, Canada

[4]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada

Correspondence
Cynthia A. Jackevicius, Western University of Health Sciences, College of Pharmacy, 309 E. Second St., Pomona, CA 91766, USA.
Email: cjackevicius@westernu.edu

Abstract

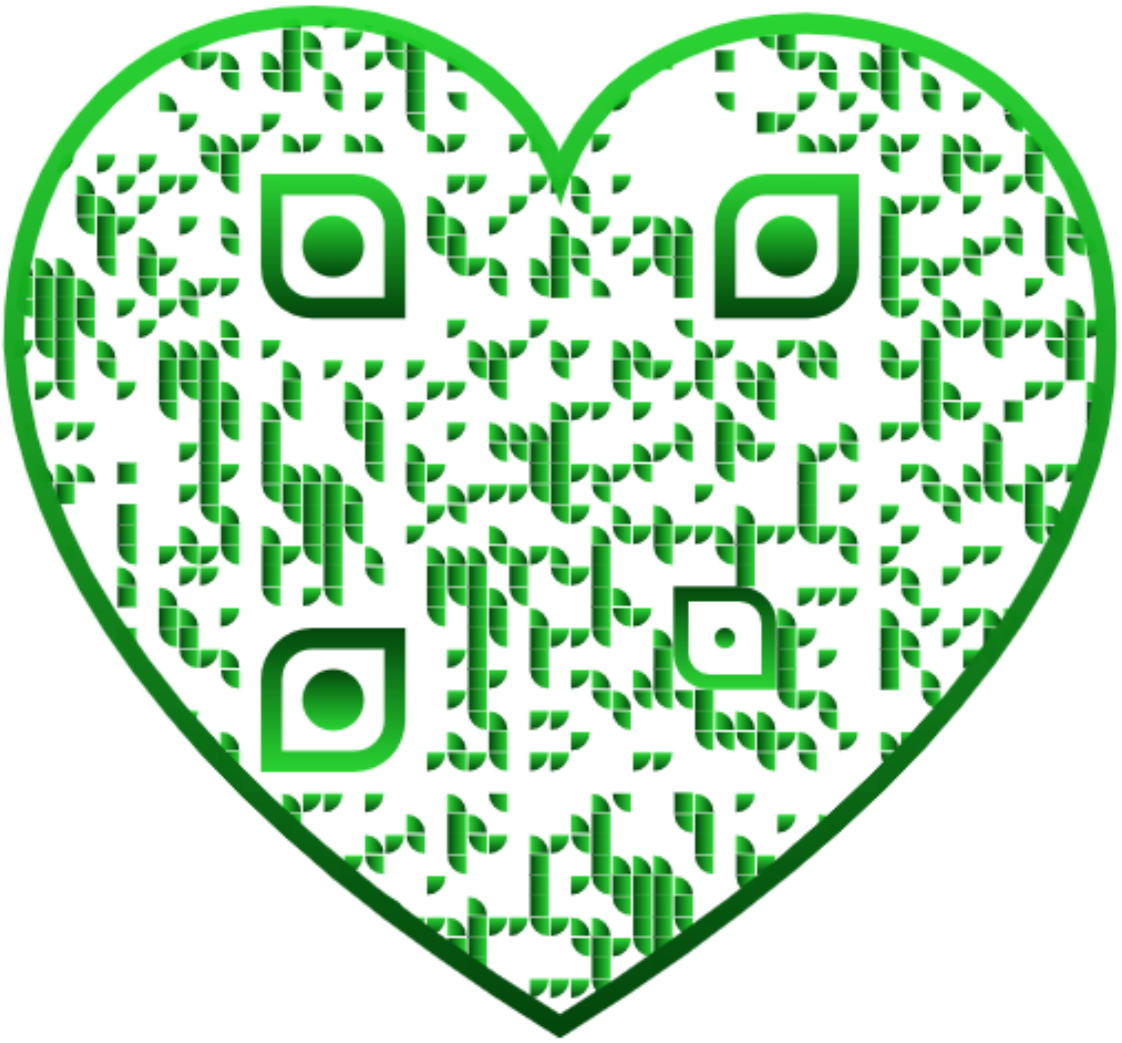
Introduction: The expanding use of Chat Generative Pre-Trained Transformer (ChatGPT, OpenAI, San Francisco, CA) for drug information may enhance access to information. However, it is crucial to assess the accuracy and reproducibility of ChatGPT responses to drug information questions, examining its utility and limitations in clinical decision-making.

Objective: To evaluate the accuracy and reproducibility of ChatGPT-3.5 and ChatGPT-4 in responding to clinician drug information questions compared with a commonly accepted resource, Lexicomp®(Wolters Kluwer Health, Philadelphia, PA).

Methods: A serial cross-sectional study was conducted on ChatGPT from March 5 to 12, 2024 in the United States. ChatGPT-3.5 is a free, artificial intelligence (AI) chatbot trained up to January 2022; ChatGPT-4 is a paid-subscription AI chatbot with internet access and more data. For trial 1 (day 0) we input 30 real-world questions (10 drug information categories) into both ChatGPT-3.5 and ChatGPT-4. For trial 2 (day 1) and 3 (day 7), 10 randomly selected questions were re-input into ChatGPT. The primary outcome evaluated the accuracy of ChatGPT-3.5 responses versus (vs.) Lexicomp® using a 4-point Likert scale. Secondary outcomes included assessing the accuracy of ChatGPT-4 responses vs. Lexicomp, comparing the accuracy of both ChatGPT versions' responses, and comparing reproducibility of ChatGPT responses over time. Cohen's Kappa and Cochran's Q assessed reproducibility.

Results: ChatGPT-3.5 demonstrated 30% accuracy (9/30), while ChatGPT-4 had 40% (12/30) accuracy ($p = 0.51$). Neither ChatGPT versions accurately answered all the questions in any category. ChatGPT-3.5's agreement between trials 1 vs. 2, 1 vs. 3, and 2 vs. 3 had fair ($k = 0.21$), moderate ($k = 0.41$), and substantial agreement ($k = 0.62$), respectively. ChatGPT-4 trials 1 vs. 2, 1 vs. 3, and 2 vs. 3 had fair

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2025 The Author(s). JACCP: *Journal of the American College of Clinical Pharmacy* published by Wiley Periodicals LLC on behalf of ACCP Foundation, Ltd.



($k = 0.23$), substantial ($k = 0.80$), and fair agreement (0.40). The accuracy of ChatGPT-3.5 vs. ChatGPT-4 for the 10 questions across the three trials was 30%, 20%, and 10% ($p = 0.78$), and 60%, 40%, and 50% ($p = 0.82$).

Conclusions: Both ChatGPT versions demonstrated limited accuracy and reproducibility in answering drug information questions, suggesting that health care professionals should exercise caution when using ChatGPT for drug information.

KEYWORDS
artificial intelligence, decision making, drug information services, reproducibility of results

1 | INTRODUCTION

The use of artificial intelligence (AI) tools offers a promising alternative to conventional subscription-based resources, especially for clinicians seeking efficient and reliable answers to drug information questions. However, there is limited data on the accuracy, reliability, and safety of drug information retrieved through AI tools. There are potential risks to a patient's health if these tools provide erroneous or misleading data, and it is uncertain which types of questions that AI tools can effectively and safely address. Despite the uncertainties, the appeal of leveraging AI tools stems from the potential to expand access to information that may be used in clinical decision-making. There has been a recent rise in the use of AI chatbots, such as Chat Generative Pre-Trained Transformer (ChatGPT, OpenAI, San Francisco, CA) that can engage in conversations and answer questions in human-like text.[1,2] Recent studies demonstrate the growing acceptance of AI[3] and ChatGPT within health care,[4] with almost half of the pharmacist respondents willing to incorporate ChatGPT into their practice.[5]

Despite the growing interest in AI, studies on ChatGPT's performance in addressing drug information questions remain limited in scope or show conflicting findings. While some studies have focused on specific questions, such as drug–drug interactions[6–8] or disease-state interactions,[9] few have explored its use in addressing a broad spectrum of practical, real-world drug information questions that are often encountered in clinical practice. Moreover, the comparison between different versions of ChatGPT, particularly between ChatGPT-3.5 (free) and ChatGPT-4 (subscription-based), has had limited evaluation.[10,11] Our study seeks to bridge the existing gaps by examining the accuracy and reproducibility of each version of ChatGPT by comparing it with a commonly accepted drug information resource.

2 | METHODS

2.1 | Study design and data sources

This was a serial cross-sectional study evaluating the use of ChatGPT software from March 5, 2024 to March 12, 2024. Two software versions were used: ChatGPT-3.5 and ChatGPT-4. ChatGPT is a generative AI chatbot based on a large language model (LLM) created by OpenAI.[12] ChatGPT-3.5, released in November 2022, is publicly available for free, whereas ChatGPT-4, released in March 2023, requires users to pay a $20/month subscription. ChatGPT-4 is a multimodal LLM with more features than ChatGPT-3.5, including the ability to create AI-generated images and internet access; however, it limits users to 30 messages per hour. The commonly accepted resource used to compare with ChatGPT's responses was Lexicomp®(Wolters Kluwer Health, Philadelphia, PA), also known as UpToDate® LexiDrug™ since March 13, 2024.[13] Lexicomp is a paid subscription-based database offering evidence-based drug information. Subscription costs vary from $70/year to $799/year depending on the type of access and user. Institutional subscriptions are also available. Lexicomp was set as the commonly accepted resource over other drug information resources such as Micromedex® (Merative, Ann Arbor, MI) due to its more frequent use, ease of navigation, and concise information.[14]

2.2 | Drug information question framework

We identified 10 drug information question categories based on clinical experience and prior research related to common question types encountered by pharmacists.[15,16] Based on real-world questions asked to or by pharmacists in the last year, we crafted three questions per category, all phrased in the perspective of a pharmacist (rather than a patient), for a total of 30 questions. At least one question per category was cardiology-focused as cardiovascular disease is prevalent and commonly managed with medications with important morbidity and mortality consequences (Table S1).[17]

The commonly accepted resource, Lexicomp, was used to compare ChatGPT responses. All answers to each question were searched in Lexicomp by two authors to create the commonly accepted resource's answers to each question prior to starting the trials with ChatGPT. The Lexicomp answer and word count were documented. Thereafter, all 30 questions were individually input into both ChatGPT-3.5 and ChatGPT-4 for trial 1 (Day 0, March 5, 2024). To prevent potential bias from using ChatGPT repeatedly, each question was input in a new chatbox after clearing the history of the previous question. Questions were input without an iterative process, meaning no follow-up questions were asked.

To evaluate the reproducibility of ChatGPT's responses over time, trials 2 (day 1) and 3 (day 7) were conducted. These two trials were

25749870, 0, Downloaded from https://accpjournals.onlinelibrary.wiley.com/doi/10.1002/jac5.70038 by Yemen Hinari NPL, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

conducted in the same manner as trial 1 except that only 10 questions were included from the original 30 questions due to the feasibility of our study. A random number generator was used to select one question from each drug information question category to use for trials 2 and 3. A total of 10 questions were repeated for the last two trials, with all 10 questions asked at three time points, each for ChatGPT-3.5 and ChatGPT-4. For trials 2 and 3, the accuracy of these 10 questions was evaluated in the same manner as trial 1. Each response and its word count for both ChatGPT versions and all three trials were documented in its entirety.

2.3 | Outcomes

Our primary outcome was the accuracy of ChatGPT responses in comparison to the commonly accepted resource, Lexicomp. The accuracy of ChatGPT responses was assessed using a four-point Likert scale (4 = completely accurate, 3 = almost accurate, 2 = partially accurate, 1 = inaccurate). Two assessors, clinical pharmacists, independently rated each ChatGPT response using the Likert scale, under the assumption that the ChatGPT response would need to encompass all the necessary information to answer the drug information question when they determined the accuracy of the response. This assumption was to minimize biases related to using pre-existing clinical knowledge when assessing the responses. Each question's ratings were evaluated for consistency between the two assessors. Discrepancies were resolved by a third, more experienced pharmacist, who made the final rating decision. Before the study, the assessors were calibrated with four practice questions followed by a discussion of any discrepancies.

Our secondary outcomes were reproducibility and usability. Reproducibility was defined as consistency in accuracy at three time points (day 0, day 1, day 7). At the end of data collection, the two assessors were asked to complete a five-question survey on their perspective of ChatGPT's responses and usefulness. The survey questions were rated using a five-point Likert scale with one being the lowest (strongly disagree, inadequate, very inconsistent) and five being the highest (strongly agree, excessive, and very consistent).[18]

2.4 | Statistical analysis

Each accuracy response was rated using a four-point Likert scale. The four-point scale was then transformed into a binary variable for statistical analysis. "Accurate" was defined as a score of 3 or 4 and "inaccurate" was defined as a score of 1 or 2. The proportion of accuracy was estimated by tallying the number of responses that were deemed accurate divided by the total number of responses, 30 for trial 1 and 10 for trials 2 and 3. McNemar's test was used to compare the proportions of accurate responses between ChatGPT-3.5 and ChatGPT-4. For reproducibility, we conducted two statistical tests: Cohen's Kappa examined the agreement between paired trials, and Cochran's Q test compared the accuracy of ChatGPT over time. Chi-square was used to compare accuracy between drug information categories. A subgroup analysis of the cardiology-related versus the non-cardiology-related drug information question was also conducted using Chi-square. One-way ANOVA was used to compare the mean word count and the Kruskal–Wallis test for median word count between Lexicomp, ChatGPT-3.5, and ChatGPT-4. A p-value of <0.05 was considered statistically significant. IBM SPSS Statistics 29.0.2.0 (Armonk, NY) was used for all statistical analyses.

3 | RESULTS

The primary results for ChatGPT-3.5 demonstrated 30% accuracy (9/30) for trial 1. ChatGPT-3.5 did not accurately answer all three questions in any of the drug information categories; however, it accurately answered two out of three questions in three drug information categories (contraindications/warnings/precautions, dosage adjustments, and pregnancy/lactation) ($p = 0.28$) (Figure 1). ChatGPT 4 demonstrated 40% accuracy (12/30). Similarly, no drug information categories had accurate responses for all three questions; however, four drug information categories demonstrated accuracy in two out of three questions (dosing, pharmacokinetics, pregnancy/lactation, and side effects/adverse drug effects) ($p = 0.56$) (Figure 1). Overall, both ChatGPT-3.5 and ChatGPT-4 answered six questions accurately and 15 questions inaccurately ($p = 0.51$) (Table 1).

ChatGPT-3.5 had 0% (0/3) accuracy for four categories including administration/preparation, drug interactions, intravenous (IV) compatibility, and monitoring parameters. ChatGPT-4 had 0% (0/3) accuracy in two categories including administration/preparation and dosage adjustments. Notably, both versions were completely inaccurate in the administration/preparation category in trial 1. One inaccurate response occurred when ChatGPT 4 identified Arexvy®(GlaxoSmithKline Biologicals, Rixensart, Belgium)as a human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS) medication instead of a vaccine for respiratory syncytial virus,[19] and ChatGPT-3.5 was not aware of this drug. Full responses can be found in Table S2. Another inaccurate response involved ChatGPT-3.5 advising that esomeprazole capsules can be crushed for nasogastric tube administration, which is incorrect.

With respect to reproducibility, ChatGPT-3.5 had accurate responses in 30%, 20%, and 10% of the 10 questions across days 0, 1, and 7, respectively, with no difference in accuracy across the trials ($p = 0.78$). ChatGPT-3.5's accuracy was consistent across the three trials for only one question related to side effects/adverse drug reactions, while inaccurate responses were consistent across the three trials for six questions: administration/preparation, dosage adjustments, dosing, drug interactions, monitoring parameters, and pharmacokinetics, with inconsistency for three questions between trials (Figure 2). ChatGPT 4's reproducibility across days 0, 1, and 7 demonstrated accuracy at a rate of 60%, 40%, and 50% ($n = 10$), respectively, with no difference in accuracy across the trials ($p = 0.82$). Accuracy across the three trials was consistent for three questions: IV compatibility, monitoring parameters, and pharmacokinetics, while inaccuracy was consistent across the three trials for three questions: administration/

25749870, 0, Downloaded from https://accpjournals.onlinelibrary.wiley.com/doi/10.1002/jac5.70038 by Yemen Hinari NPL, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

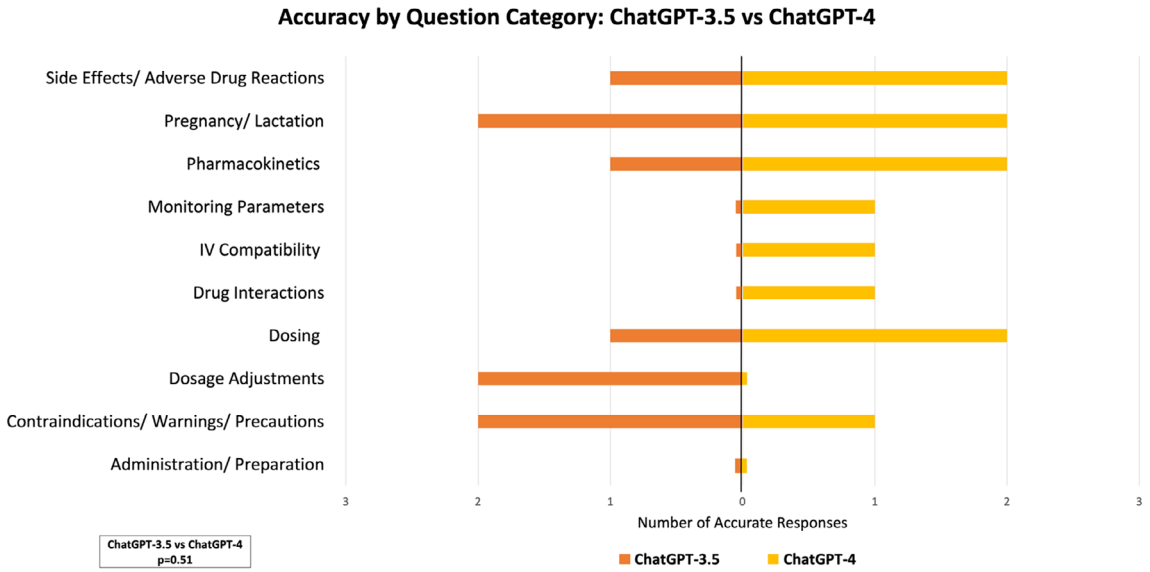


FIGURE 1 Number of accurate responses for ChatGPT-3.5 vs. ChatGPT-4 in each of the 10 drug information categories for Trial 1 (Day 0). The Y-axis shows 10 drug information question categories and the X-axis shows the number of accurate responses out of 3, bidirectionally. ChatGPT-3.5 accuracy is on the left and ChatGPT-4 accuracy is on the right. No differences were found in the accuracy of responses for ChatGPT-3.5 vs. ChatGPT-4 ($p = 0.51$). IV = intravenous.

TABLE 1 Key performance metrics for ChatGPT-3.5 and ChatGPT-4.

Metric		ChatGPT-3.5	ChatGPT-4	*p* value	Notes
Overall accuracy[a]		30% (9/30)	40% (12/30)	0.51[b]	Accuracy defined as responses rated 3 or 4 on the Likert scale
Accuracy by category		0%–67% across categories	0%–67% across categories	N/A	Both versions performed better in the pregnancy/ lactation category and worse in others
Reproducibility (Kappa)[c]		0.21–0.62 (fair–moderate)	0.23–0.80 (fair–substantial)	N/A	Agreement across repeated trials (Day 0, 1, 7)
Reproducibility over time		30%, 20%, 10%	60%, 40%, 50%	N/A	% accuracy across days for repeated questions
Word count[a]	Mean ± SD[d]	153.6 ± 82.3	201.1 ± 110.0	<0.001	ChatGPT-4 provided more detailed, longer responses compared with ChatGPT-3.5
	Median[e]	139.5	215	0.035	

[a]From Trial 1.
[b]*p*-value reflects ChatGPT-3.5 vs. ChatGPT-4 comparison.
[c]Kappa values measure agreement between repeated trials: 0.0–0.20 (slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial).
[d]Reference (Lexicomp): 30.2 ± 29.2.
[e]Reference (Lexicomp): 23.0.

preparation, dosage adjustments, and dosing, with inconsistency for four questions between trials (Figure 2). Table S3 illustrates an example of inconsistency in responses over time. When examining the agreement between paired trials, we found that ChatGPT-3.5 trials 1 versus (vs.) 2, 1 vs. 3, and 2 vs. 3 had fair ($k = 0.21$), moderate ($k = 0.41$), and substantial agreement ($k = 0.62$), respectively. ChatGPT-4 trials 1 vs. 2, 1 vs. 3, and 2 vs. 3 had fair ($k = 0.23$), substantial ($k = 0.80$), and fair agreement (0.40). No significant differences were observed when comparing ChatGPT-3.5 to ChatGPT-4 with the 10 questions across the three trials ($p = 0.38$, $p = 0.63$, $p = 0.13$; Table 1).

The subgroup analysis for cardiology-specific questions showed that ChatGPT-3.5 had an accuracy of 41.7% for cardiology questions ($n = 12$) and 22.2% for non-cardiology questions ($n = 18$) ($p = 0.42$). In comparison, ChatGPT-4 demonstrated an accuracy of 50% for cardiology-specific questions ($n = 12$) and 33.3% for non-cardiology questions ($n = 18$) ($p = 0.46$).

The usability and feasibility of ChatGPT's responses were assessed based on word count and a five-question assessor survey. A difference was observed in the mean word count with standard deviation between ChatGPT-3.5, ChatGPT-4, and Lexicomp (153.6 ± 82.3, 201.1 ± 110.0, 30.2 ± 29.2; $p < 0.001$). For ChatGPT-3.5, ChatGPT-4, and Lexicomp, the median word count across all questions for trial 1 was 139.5, 215.0, and 23.0 words ($p = 0.035$; Table 1). Notably, based on survey results, one assessor described the amount of information provided by ChatGPT as "excessive", with both assessors

25749870, 0, Downloaded from https://accpjournals.onlinelibrary.wiley.com/doi/10.1002/jac5.70038 by Yemen Hinari NPL, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ChatGPT-3.5 Accuracy*			Question	ChatGPT-4 Accuracy*		
Day 0	Day 1	Day 7		Day 0	Day 1	Day 7
🔴	🟢	🔴	Are calcium gluconate and ceftriaxone Y-site compatible?	🟢	🟢	🟢
🟢	🔴	🔴	Can a breastfeeding patient start ethinyl estradiol and norgestimate birth control two months postpartum?	🟢	🔴	🔴
🔴	🔴	🔴	How is esomeprazole capsule administered in a patient with an NG tube?	🔴	🔴	🔴
🔴	🔴	🔴	How is morphine renally adjusted in a patient with creatinine clearance of 25 ml/min?	🔴	🔴	🔴
🔴	🔴	🔴	How often do we need to monitor LFTs for amiodarone?	🟢	🟢	🟢
🟢	🟢	🟢	Is hydrochlorothiazide associated with skin cancer?	🟢	🔴	🟢
🟢	🔴	🔴	Is it okay to use haloperidol in a patient with Parkinson's disease?	🔴	🟢	🔴
🔴	🔴	🔴	What are the drug interactions with Paxlovid, diltiazem, and lisinopril?	🟢	🔴	🟢
🔴	🔴	🔴	What is the half-life of daridorexant?	🟢	🟢	🟢
🔴	🔴	🔴	What is the starting dose of alirocumab?	🔴	🔴	🔴
30%	20% (p=0.78)	10%	Total Accurate % (p-value)	60%	40% (p=0.82)	50%

FIGURE 2 Reproducibility in accuracy over time for ChatGPT-3.5 and ChatGPT-4 at Days 0, 1, and 7. Ten randomly selected questions (one from each drug information category) are displayed in the middle of the figure. Accurate responses are represented with green bullet and inaccurate responses are represented with red bullet. On the left is ChatGPT-3.5's accuracy for days 0, 1, and 7 with total percentage accurate and p-value at the bottom of each column. On the right is ChatGPT-4's accuracy for days 0, 1, and 7 with total percentage accurate and *p*-value at the bottom of each column.

indicating that they would not recommend using ChatGPT to a health care provider. Additional survey results are summarized in Table S4.

4 | DISCUSSION

We found that both ChatGPT-3.5 and ChatGPT-4 had poor accuracy when answering real-world, practical drug information questions, with less than half of the responses deemed accurate in the primary analysis. The reproducibility analyses demonstrated fair to moderate agreement in most of the trial comparisons. Usability was rated low based on the word count and the assessor's survey findings. Our comprehensive study assessed accuracy and reproducibility with a wide range of practical drug information questions for both versions of ChatGPT, unlike previous studies that only focused on one version or one question category.

Our study demonstrated both ChatGPT versions had limited accuracy for broad-based, real-world drug information questions, with ChatGPT-3.5 inaccurately answering all three questions in four of the 10 categories. Some other studies also found limited accuracy when examining ChatGPT. For example, one study of 50 drug-related questions across varied categories found that ChatGPT-3.5 had 26% completely correct and 36% partially correct, despite using different commonly accepted resources.[20] Another study asking questions from a health care professionals' perspective found that only 25% of ChatGPT-3.5's responses were accurate.[21]

While our study found no statistical differences in the accuracy of the responses between categories of drug information, there were noteworthy qualitative differences in ChatGPT's performance across the drug information categories. Both ChatGPT-3.5 and ChatGPT-4 performed better, accurately answering two out of the three questions in the pregnancy/lactation category. However, both versions showed poor accuracy for the drug interactions category for both ChatGPT-3.5 and ChatGPT-4, aligning with a previous study.[7] In contrast, a study on drug interactions demonstrated 100% accuracy ($n = 39$) with ChatGPT-4, but accuracy decreased with more complex cases. This discrepancy may be due to the study using a 70% threshold for accuracy.[6] Our exploratory subgroup analysis found that cardiology-specific questions numerically had a higher accuracy rate compared with non-cardiology drug information questions, suggesting a need for further research based on therapeutic area.

Our study, along with others,[10] demonstrated that ChatGPT-3.5 and ChatGPT-4 exhibited limited accuracy, consistent with findings for other AI chatbots. For example, one study compared drug–drug interaction accuracy among AI tools and found Microsoft Bing AI (Microsoft Corporation, Redmond, WA) to be the most accurate (89%), followed by Google Bard (Google LLC, Mountain View, CA) (68.6%), while ChatGPT-3.5 and ChatGPT-4 showed lower accuracy rates (52.5% and 59.2%, respectively).[8] Munir and colleagues reported that ChatGPT performs well in drug information and calculation tasks but struggles with complex patient cases.[22] Another study compared ChatGPT's responses to those of clinical pharmacists, finding strengths in drug counseling but significant weaknesses in advanced reasoning, such as prescription review.[23] These differences suggest that certain AI models may provide more reliable responses for specific clinical applications. Other AI models, including Microsoft Bing AI[8] or Google Bard[24] may offer better accuracy. Grossman and colleagues highlighted ChatGPT's tendency to provide incomplete responses.[16] Similarly, another study found ChatGPT's responses to be more readable than those of traditional drug information centers but less accurate, especially when referencing reliable sources.[25] Our study focused on ChatGPT models due to their recent popularity compared with the other AI models. However, tools like Claude-2 Claude-2 (Anthropic, San Francisco, CA), which outperformed ChatGPT-3.5 in medical physiology questions (72.7% vs. 47.3%), highlight the need for further investigation for drug information.[11] These

25749870, 0, Downloaded from https://accpjournals.onlinelibrary.wiley.com/doi/10.1002/jac5.70038 by Yemen Hinari NPL, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

findings emphasize the importance of selecting AI tools based on the type of clinical information needed and their demonstrated accuracy.

Potential reasons for the limited accuracy of ChatGPT in our study include AI hallucinations, lack of an iterative process, and outdated data. AI hallucinations pose a serious concern, as they can fabricate responses to produce false or misleading information.[26–29] This phenomenon is particularly noted in ChatGPT-4 and other AI systems that produce text, images, or other media. An example occurred when ChatGPT-4 incorrectly responded, "Arexvy, also known by its generic name ibalizumab-uiyk, is a medication used for the treatment of HIV/AIDS". Although, ChatGPT-4 has internet access, this response highlights concern about whether ChatGPT can be a trusted resource, especially when considering patient safety. Research on the impact or frequency of AI hallucinations is limited, but newer models, including Claude-3, claim lower AI hallucination rates[30] which requires verification in practice. As AI models evolve, health care providers must remain cautious of their potential in disseminating false information that can result in harmful patient outcomes. For both versions of ChatGPT, the lack of using an iterative process (not asking follow-up questions) may have reduced accuracy, however, an iterative requirement would decrease the convenience of using an AI chatbot. Additionally, ChatGPT-3.5 has only been trained on data up to January 2022, which may account for some inaccuracies, especially with newer medications.

Our study found a lack of reproducibility in the accuracy of responses. For example, when asking ChatGPT-3.5 about Y-site compatibility of ceftriaxone and calcium gluconate, trial 1, 2, and 3 responses were deemed inaccurate, accurate, and inaccurate, respectively. This demonstrates that in as little as one day, ChatGPT-3.5 changed its response, which indicates that depending on the given day, a provider may get inaccurate information from ChatGPT, compromising patient safety. The lack of consistent delivery of accurate evidence-based information demonstrates deficiencies in current AI models and demonstrates the need for clinicians to exercise caution when using such tools.

Prior studies on ChatGPT's reproducibility showed varying results. One study found low reproducibility with ChatGPT-3.5, when two of the three repeated questions had varying responses over several weeks.[20] Another study testing ChatGPT-3.5 showed 79.2% agreement between responses in weeks 1 and 3, and 87.5% between weeks 3 and 5.[31] A study of both ChatGPT-3.5 and 4 found a 65% agreement level ($K = 0.394$, fair agreement) 2 days apart for ChatGPT-3.5, and 80% ($K = 0.467$, moderate agreement) for ChatGPT-4.[10] These results align with our study as most of our reproducibility results indicated fair to moderate agreement. Our rationale for conducting our study on days 0, 1, and 7 was to be able to determine whether any differences would be seen in as little as 1 day, but also up to 1 week later.

Our study found ChatGPT's usability to be low with objective and subjective measures. Both ChatGPT versions had a significantly higher word count compared with Lexicomp, with ChatGPT-4 having the numerically highest mean and median word count. ChatGPT responses used at least five times more words when compared with Lexicomp. The high word burden may be time-consuming, overwhelming, and make finding relevant information within a ChatGPT response difficult, which can potentially contribute to medical errors.[32] Additionally, studies have identified that longer sentences increase cognitive load and reduce comprehension efficiency.[33–37] Our assessors rated the amount of information provided by ChatGPT as "excessive" or at best, neutral, with one assessor indicating that they "disagreed" when asked if the relevant information was easy to identify within the ChatGPT responses, further demonstrating the inefficiency of locating the appropriate information within each response. Furthermore, both assessors would not recommend using ChatGPT to a health care provider. To our knowledge, no other studies have evaluated the usability or the word burden of ChatGPT.

Our study has some limitations. First, our study may have been underpowered to detect statistically significant differences between question categories given our small sample size of questions. However, because we assessed both versions of ChatGPT, having 30 questions in trial 1 and 10 questions for the reproducibility portion of our study was found to be most feasible for our study. Second, the four-point Likert scale used for rating each response required subjectivity, although when the assessors were calibrated with sample questions, 100% agreement was achieved prior to data collection. Third, a five-point Likert scale was used to survey our assessors on their perspective of ChatGPT responses and its usability, which allowed the assessors to choose a neutral stance, potentially biasing towards the null. Fourth, although best practice in answering drug information questions involves using more than one resource to ensure accuracy and comprehensive information, to ensure consistency and reproducibility of answers, our study used one commonly accepted resource. Finally, further studies are warranted to address specific therapeutic areas, newer AI models, and usability of AI chatbots.

5 | CONCLUSION

Our study highlights the inaccuracy of ChatGPT-3.5 and ChatGPT-4 in answering real-world drug information questions. The use of ChatGPT and other forms of AI may compromise patient safety, especially in instances of AI hallucinations. Dissemination of inaccurate information not only burdens health care providers with added workload, but also jeopardizes patient well-being by influencing clinical decisions. While the findings of our study suggest ChatGPT has poor accuracy when answering drug information questions, future studies with larger sample sizes and broader drug information questions may be warranted to further investigate and validate the accuracy and clinical utility of AI chatbots within our health care system. Given the technological advances and constant evolvement of AI tools, health care providers should exercise caution when using ChatGPT for drug information as compared with evidence-based resources.

FUNDING INFORMATION

There was no external funding for this study.

CONFLICT OF INTEREST STATEMENT

Dr. Jackevicius is a member of the JACCP editorial board. All other authors declare no conflicts of interest.

25749870, 0, Downloaded from https://accpjournals.onlinelibrary.wiley.com/doi/10.1002/jac5.70038 by Yemen Hinari NPL, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Shikha Khatri https://orcid.org/0009-0000-7237-7040
Anthony Sengul https://orcid.org/0009-0009-1641-9387
Jungyeon Moon https://orcid.org/0000-0002-7443-6953
Cynthia A. Jackevicius https://orcid.org/0000-0001-5274-7251

REFERENCES

1. Nerdynav. 107 Up-to-Date ChatGPT statistics & user numbers [April 2024]. 2024 https://nerdynav.com/chatgpt-statistics/
2. Notta. 100+ incredible ChatGPT statistics & facts in 2024. 2024 https://www.notta.ai/en/blog/chatgpt-statistics
3. Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: opportunities and challenges. *Int J Environ Res Public Health*. 2021;18(1):271.
4. Chow MS. Impact of artificial intelligence: a perspective on human life and pharmacy practice. *J Am Coll Clin Pharm*. 2023;6(3):216–219. https://doi.org/10.1002/jac5.1705
5. Abu-Farha R, Fino L, Al-Ashwal FY, Zawiah M, Gharaibeh L, Harahsheh MM, et al. Evaluation of community pharmacists' perception and willingness to integrate ChatGPT into their pharmacy practice: a study from Jordan. *J Am Pharm Assoc*. 2023;63(6):1761–1767.
6. Roosan D, Padua P, Khan R, Hasiba K, Verzosa C, Wu Y. Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management. *J Am Pharm Assoc*. 2023;64(2):422–428.
7. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The capability of ChatGPT in predicting and explaining common drug-drug interactions. *Cureus*. 2023;15(3):e36272.
8. Al-Ashwal FY, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar AN. Evaluating the sensitivity, specificity, and accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and bard against conventional drug-drug interaction clinical tools. *Drug Healthc Patient Saf*. 2023;15:137–147.
9. Gao Z, Li L, Ma S, Wang Q, Hemphill L, Xu R. Examining the potential of ChatGPT on biomedical information retrieval: fact-checking drug-disease associations. *Ann Biomed Eng*. 2023 2023, Apr 17;52(8):1919–1927. https://doi.org/10.1007/s10439-023-03385-w
10. Bazzari FH, Bazzari AH. Utilizing ChatGPT in telepharmacy. *Cureus*. 2024;16(1):e52365. https://doi.org/10.7759/cureus.52365
11. Agarwal M, Goswami A, Sharma P. Evaluating ChatGPT-3.5 and Claude-2 in answering and explaining conceptual medical physiology multiple-choice questions. *Cureus*. 2023;15(9):e46222.
12. Open AI. Ask ChatGPT anything. 2024 https://openai.com/
13. Wolters Kluwer. UpToDate Lexidrug: Evidence-based drug referential content for teams. 2024 https://www.wolterskluwer.com/en/solutions/uptodate/enterprise/lexidrug#form
14. Mountford CM, Lee T, Lemos J, Loewen PS. Quality and usability of common drug information databases. *Can J Hosp Pharm*. 2010;63(2):130–137.
15. Hou CL, Lu YH, Chien SC, Chen HH, Chen CY. Questions frequently asked of healthcare professionals: a 2-year data survey conducted at a medical center. *Biomed Res Int*. 2018;18:4895850.
16. Grossman S, Zerilli T, Nathan JP. Appropriateness of ChatGPT as a resource for medication-related questions. *Br J Clin Pharmacol*. 2024;90(10):2691–2695.
17. Centers for Disease Control and Prevention. About Chronic Diseases. 2024 https://www.cdc.gov/chronic-disease/about/index.html
18. Jackevicius CA, Lash D, Singh D, Hines K, Hata M. Accuracy and usability of medication identifiers for solid oral medications. *J Am Coll Clin Pharm*. 2018;1(2):103–109.
19. Arexvy. Arexvy [package insert on the internet]. Belgium: GlaxoSmithKline Biologicals. 2024 https://www.fda.gov/media/167805/download
20. Morath B, Chiriac U, Jaszkowski E, Deiß C, Nürnberg H, Hörth K, et al. Performance and risks of ChatGPT used in drug information: an exploratory real-world analysis. *Eur J Hosp Pharm*. 2023;31(6):491–497. https://doi.org/10.1136/ejhpharm-2023-003750
21. Hsu HY, Hsu KC, Hou SY, Wu CL, Hsieh YW, Cheng YD. Examining real-world medication consultations and drug-herb interactions: ChatGPT performance evaluation. *JMIR Med Educ*. 2023;9:e48433.
22. Munir F, Gehres A, Wai D, Song L. Evaluation of ChatGPT as a tool for answering clinical questions in pharmacy practice. *J Pharm Pract*. 2024;37(6):1303–1310.
23. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: a comparative study of ChatGPT and clinical pharmacists. *Br J Clin Pharmacol*. 2024;90(1):232–238.
24. Sulaiman DM, Shaba SS, Almufty HB, Sulaiman AM, Merza MA. Screening the drug-drug interactions between antimicrobials and other prescribed medications using Google bard and Lexicomp® online™ database. *Cureus*. 2023;15(9):e44961.
25. Triplett S, Ness Engle GL, Behnen EM. A comparison of drug information question responses by a drug information center and by ChatGPT. *Am J Health-Syst Pharm*. 2024;82(8):448–460.
26. Menz BD, Modi ND, Sorich MJ, Hopkins AM. Health disinformation use case highlighting the urgent need for artificial intelligence vigilance: weapons of mass disinformation. *JAMA Intern Med*. 2024;184(1):92–96.
27. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med*. 2023;388(13):1233–1239.
28. Hatem R, Simmons B, Thornton JE. A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. *Cureus*. 2023;15(9):e44720.
29. Ratwani RM, Bates DW, Classen DC. Patient safety and artificial intelligence in clinical care. *JAMA Health Forum*. 2024;5(2):e235514.
30. Anthropic. Meet Claude. 2024 https://www.anthropic.com/claude
31. Al-Dujaili Z, Omari S, Pillai J, al Faraj A. Assessing the accuracy and consistency of ChatGPT in clinical pharmacy management: a preliminary analysis with clinical pharmacy experts worldwide. *Res Social Adm Pharm*. 2023;19(12):1590–1594. https://doi.org/10.1016/j.sapharm.2023.08.012
32. Edmunds A, Morris A. The problem of information overload in business organisations: a review of the literature. *J Inf Manag*. 2000;20(1):17–28.
33. Vizer LM, Sears A. Efficacy of personalized models in discriminating high cognitive demand conditions using text-based interactions. *Int J Hum Comput Stud*. 2017;104:80–96.
34. Asgari E, Kaur J, Nuredini G, Balloch J, Taylor AM, Sebire N, et al. Impact of electronic health record use on cognitive load and burnout among clinicians: narrative review. *JMIR Med Inform*. 2024;12:e55499. https://doi.org/10.2196/55499
35. Mikk J. Sentence length for revealing the cognitive load reversal effect in text comprehension. *Educ Stud*. 2008;34(2):119–127.
36. Gibson E. The dependency locality theory: a distance-based theory of linguistic complexity. In: Marantz AP, Miyashita Y, O'Neil W, editors. *Image, language, brain: papers from the first mind articulation project symposium*. Cambridge, MA: MIT Press; 2001. p. 95–126.
37. Just MA, Carpenter PA. A capacity theory of comprehension: individual differences in working memory. *Psychol Rev*. 1992;99(1):122–149.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Khatri S, Sengul A, Moon J, Jackevicius CA. Accuracy and reproducibility of ChatGPT responses to real-world drug information questions. *J Am Coll Clin Pharm*. 2025;1-7. doi:10.1002/jac5.70038

25749870, 0, Downloaded from https://accpjournals.onlinelibrary.wiley.com/doi/10.1002/jac5.70038 by Yemen Hinari NPL, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License